A DEMONSTRATION OF 'REGRESSION TOWARD THE MEAN'

Occasional Research Paper, No. 17
October 2010

Edward B. Reeves, Director
Center for Educational Research & Leadership
School of Public Affairs
Morehead State University

ABSTRACT: This paper provides a brief discussion and demonstration of regression toward the mean, a subtle statistical artifact that appears in the effort to measure change. Regression toward the mean frequently arises in educational assessment when repeated testing is used to determine achievement growth among students or schools at the extremes of the achievement distribution. This statistical artifact is important because it can lead to erroneous inferences about what is causing the observed changes in test scores. The demonstration makes use of Kentucky Core Content Test data.

A Demonstration of 'Regression Toward the Mean'

This paper provides a brief discussion and demonstration of 'regression toward the mean', a statistical phenomenon that is also referred to by the following terms: 'regression to the mean', 'mean reversion', and 'regression artifacts'.[1] Regression toward the mean arises in the attempt to measure change using repeated measures. The effects of regression toward the mean are frequently noted in educational testing, although they are often incorrectly attributed to measurement error or to unique conditions affecting the test results such as outstanding teaching, student motivation, and the like. Stated simply, regression toward the mean may be present when students (or schools) that scored at the high or low ends of the test score distribution on the initial test are found to score closer to the mean on subsequent testing. Because of this, the regression artifact has a characteristic "signature": **the change score is negatively correlated with the initial test score.**

Regression toward the mean may mislead policymakers, education administrators, teachers, and others to wrongly interpret the reason for test score changes. The following examples illustrate how this can happen:

1. The state legislature authorizes a program of special assistance to very low performing schools. After a year of receiving this assistance, these schools achieve an above average improvement on their state's assessment tests. The news media report the improvement as evidence that the assistance program is working.

2. One of the state's top performing high schools finds that its test scores have declined a year later. The exasperated principal tells parents that the decline in performance can be explained by a change in school policy: a major rearrangement of teachers' schedules took place during the year. Teachers experienced difficulties adjusting to the new schedule and this caused them to be less effective in the classroom.

In both of these hypothetical examples, the reasons given for the change in test scores (highly welcomed in the first instance and a source of worry in the second) could be correct. However, regression toward the mean could also bring about the very same changes. Therefore, the explanations given for the changes in the scores are confounded by regression toward the mean.

The phenomenon was first identified by Francis Galton who noted that the children of tall parents were generally shorter as adults than their parents while the children of short parents were generally taller adults than their parents. These regression artifacts were not the result of any biological process, Galton realized, but were intrinsic to the comparison of height differences between generations. The same general phenomenon is encountered in educational assessment of achievement growth.

[1] The author wishes to thank Ben Oldham and David Sloan for their helpful comments.

If the goal is solely that of estimating the average growth of achievement for the student population as a whole, we needn't be concerned about regression toward the mean disturbing the results. The problem only arises if we select a case or a group of cases where the initial test score diverges from the mean. For example, if we want to ascertain whether a low-achieving or a high-achieving school has brought about growth in achievement, we do this by comparing the first year's test results of its students with the next year's results. But when schools are selected in this manner, regression artifacts will appear: for example, the low-achieving school may be observed to make above-average gains in its test scores while the high-achieving school is discovered making below-average gains.

Regression toward the mean may appear to be a mysterious force that drives the members of a population toward the average condition, but it is actually a well-understood statistical artifact that is a direct product of the imperfect correlation between the two sets of scores that are used to construct the change score. Campbell & Kinney (1999) show that regression toward the mean increases as the correlation coefficient (Pearson r) relating the two sets of test scores departs from unity and that its effects are most noticeable at the extremes of the distribution. Campbell & Kinney offer a mathematical proof that regression toward the mean will necessarily occur provided that the two test score distributions have similar variances and are correlated at less than unity. Because of its close relationship to the correlation coefficient, the formula for estimating the percent of regression effects is given below:

$$(1 - r)*100$$

where r is the Pearson correlation coefficient. Thus, if two tests are correlated such that r = 0.5, then the percent of regression toward the mean is 50 percent.

I have prepared a demonstration of the effects of regression toward the mean using Kentucky Core Content Test (KCCT) data from prior years. I chose to focus on 4th grade reading scale scores. (If I had chosen 5th grade math scale scores instead, the results would not have been substantially different.) I had available scores from 1999 to 2005 (see Table 1). By inspection, I settled on the medial years in the series, 2002 and 2003. These two sets of test scores (RDSS_02 and RDSS_03) had similar standard deviations and were correlated (r = 0.737) in the middle range of the Pearson r coefficients for adjacent years in the series (see the diagonal in Table 2). The average reading scale score increased from 547.97 in 2002 to 549.52 in 2003, a change or growth of 1.55 points from one spring's testing to the next (see Table 3). This growth estimate is termed the 'raw change score' because it is calculated by subtracting the initial test score from the subsequent test score (without any adjustments). It is this method of estimating change that bears the "signature" of regression toward the mean that was mentioned previously.

The scatter plot with the fit line in Figure 1 shows the effect of regression toward the mean on the raw change score. Elementary schools that scored lowest in reading in 2002 tended to have raw change scores that were larger than elementary schools that posted the

highest reading scores in 2002. Furthermore, the scatter plot and fit line show that schools with high reading scores on the 2002 test more often registered negative raw change scores than did the lesser performing schools. Table 4 and Table 5 confirm these findings in more detail, showing the raw change scores for the fifty lowest scoring elementary schools and the fifty highest scoring elementary schools respectively. The bottom panels of Tables 4 and 5 reveal the fifty lowest performing schools have a mean raw change score of +8.1 (well above the overall mean of 1.55) while the fifty highest performing schools have a mean raw change score equal to -6.1.

Campbell & Kinney (1999, Chapter 6, "Regression Artifacts in Change Scores") review several ways of correcting the errors in estimating change scores that are brought about by regression artifacts. I calculated one of these, called the 'residualized change score', to compare it with the raw change score. Campbell & Kinney note that the residualized change score has been around for decades and is well known among psychometricians. Stated intuitively, it estimates the expected change score, absent the effect of regression toward the mean. There are two main drawbacks to using the residualized change score calculation: it may not be as easily understood by the public as the raw change score, and those who benefit from the regression artifact elevating their raw change score will not like it.

The formula for calculating the residualized change score is shown below (note: Campbell & Kinney use different symbols):

$$Y_i - b_{xy}*(X_i - \bar{X}) - \bar{Y}$$

With respect to the present study, these symbols represent the following:

Y_i is the school's 2003 reading scale score.

b_{xy} is the regression coefficient obtained when the 2003 readings scores were regressed on the 2002 scores (in this case, $b_{xy} = 0.741$).

$(X_i - \bar{X})$ is the school's 2002 reading score expressed as a deviation from the mean score.

$\bar{Y}$ is the 2003 mean reading score.

The effects of using the residualized change score are shown in Figure 2 and in Tables 4 and 5. The scatter plot and fit line in Figure 2 show very clearly that using this modified form of the change score eliminates the negative correlation seen in Figure 1. Table 4 shows that the residualized change scores moderate the big raw score gains seen in some of the lowest scoring schools. Meanwhile, low scoring schools that had a negative raw change score receive an even more negative residualized change score, after regression toward the mean has been controlled. In Table 5, the residualized change scores show the opposite results. The low raw change scores of high performing schools are elevated

by the residualized change score adjustment, and some instances of negative growth are reversed in this top scoring group of schools.

It is important to note that the formula for residualized change scores cited above employs a simple linear estimate of regression toward the mean. This calculation assumes that regression toward the mean has a constant effect across the distribution. This may be an unwarranted assumption in some situations. Chay, McEwan, and Urquiola (2005) have recommended the use of a cubic polynomial function, instead of the linear function, when accounting for regression to the mean. The cubic function models the assumption that regression artifacts are disproportionately larger in the tails of the distribution than they are closer to the mean. In other words, the effects of regression toward the mean are modeled to be non-linear. I explored if this method would produce substantially different results from those reported above, which used the linear function. In this instance, there were negligible differences in the results obtained from using the two methods.

The conclusion to be drawn from this brief discussion and demonstration is that regression toward the mean may well exert an influence when test score change (or achievement growth) is used as an assessment criterion. The likelihood this will happen increases when the correlation between the test scores departs substantially from unity. Then, if precautions to minimize regression artifacts are not taken, erroneous inferences about what test score changes mean for low and high performing schools may result. And such erroneous inferences may lead to the implementation of policies, including penalties and rewards for teachers and for schools, that have unintended consequences.

References

Campbell, D. T. & Kenney, D. A. (1999). *A primer on regression artifacts.* New York: Guilford Press.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review, 95*, 1237-1258.

Table 1. Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
RDSS_99	733	495	596	543.73	14.516
RDSS_00	735	504	586	545.02	13.784
RDSS_01	732	503	593	546.53	13.009
RDSS_02	735	508	595	547.97	13.400
RDSS_03	735	511	608	549.52	13.485
RDSS_04	727	515	596	553.69	13.078
RDSS_05	725	517	606	554.20	13.860
Valid N (listwise)	721				

Table 2. Correlations

		RDSS_99	RDSS_00	RDSS_01	RDSS_02	RDSS_03	RDSS_04	RDSS_05
RDSS_99	Pearson Correlation	1	.739**	.693**	.642**	.583**	.532**	.495**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	733	733	731	733	733	727	723
RDSS_00	Pearson Correlation	.739**	1	.718**	.675**	.595**	.543**	.513**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000
	N	733	735	732	735	735	727	725
RDSS_01	Pearson Correlation	.693**	.718**	1	.740**	.680**	.606**	.595**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000
	N	731	732	732	732	732	725	722
RDSS_02	Pearson Correlation	.642**	.675**	.740**	1	.737**	.678**	.622**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000
	N	733	735	732	735	735	727	725
RDSS_03	Pearson Correlation	.583**	.595**	.680**	.737**	1	.748**	.677**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000
	N	733	735	732	735	735	727	725
RDSS_04	Pearson Correlation	.532**	.543**	.606**	.678**	.748**	1	.752**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000
	N	727	727	725	727	727	727	723
RDSS_05	Pearson Correlation	.495**	.513**	.595**	.622**	.677**	.752**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	
	N	723	725	722	725	725	723	725

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3. Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
RDSS_02	735	508	595	547.97	13.400
RDSS_03	735	511	608	549.52	13.485
Raw Change Score	735	-31.00	42.00	1.5510	9.75377
Residualized Change Score	735	-31.51	45.12	-.0003	9.11753
Valid N (listwise)	735				

Fig. 1 Scatter Plot with Fit Line Relating Raw Change Score and Reading Scale Score 2002

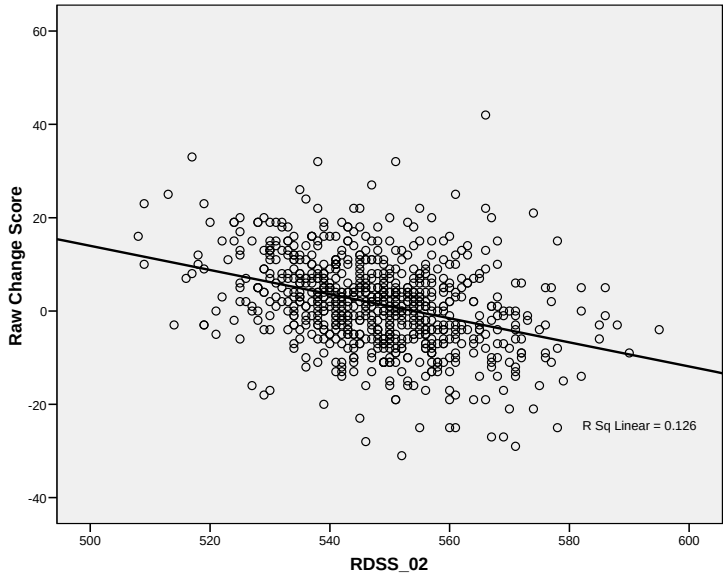

Fig. 2 Scatter Plot with Fit Line Relating Residualized Change Score and Reading Scale Score 2002

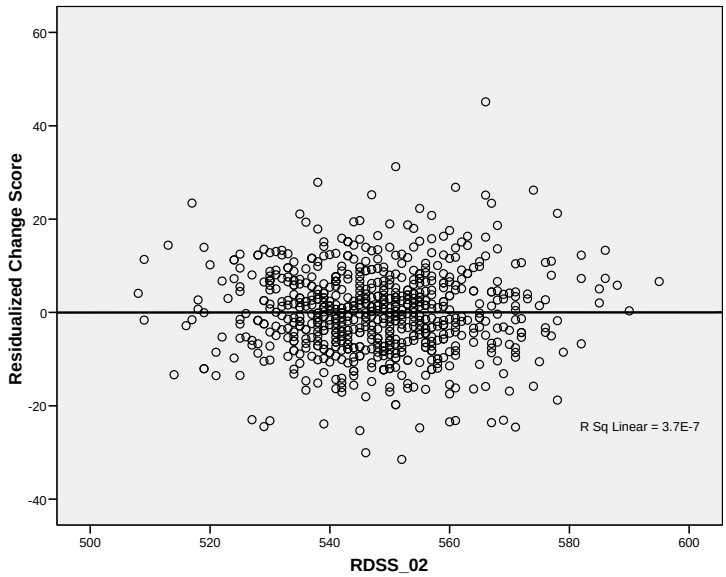

Table 4. 50 Lowest Scoring Elementary Schools, 4th Grade Reading 2002

Case #	RDSS_02	RDSS_03	Raw Chg Score	Residualized Chg Score
1	508	524	16	4.1
2	509	519	10	-1.7
3	509	532	23	11.4
4	513	538	25	14.4
5	514	511	-3	-13.4
6	516	523	7	-2.8
7	517	525	8	-1.6
8	517	550	33	23.4
9	518	528	10	0.7
10	518	530	12	2.7
11	519	516	-3	-12.1
12	519	516	-3	-12.1
13	519	528	9	-0.1
14	519	542	23	13.9
15	520	539	19	10.2
16	521	516	-5	-13.5
17	521	521	0	-8.5
18	522	525	3	-5.3
19	522	537	15	6.7
20	523	534	11	3.0
21	524	522	-2	-9.8
22	524	539	15	7.2
23	524	543	19	11.2
24	524	543	19	11.2
25	525	519	-6	-13.5
26	525	527	2	-5.5
27	525	530	5	-2.5
28	525	531	6	-1.5
29	525	537	12	4.5
30	525	538	13	5.5
31	525	542	17	9.5
32	525	545	20	12.5
33	526	528	2	-5.2
34	526	533	7	-0.2
35	527	511	-16	-23.0
36	527	527	0	-7.0
37	527	528	1	-6.0
38	527	542	15	8.0
39	528	526	-2	-8.7
40	528	528	0	-6.7
41	528	533	5	-1.7
42	528	533	5	-1.7
43	528	547	19	12.3
44	528	547	19	12.3
45	529	511	-18	-24.5
46	529	525	-4	-10.5
47	529	533	4	-2.5
48	529	533	4	-2.5

49	529	542	13	6.5
50	529	549	20	13.5
Mean	522.8	530.9	8.1	0.0
SD	5.6	10.3	10.5	10.1
Min	508	511	-18	-24.5
Max	529	550	33	23.4

Table 5. 50 Highest Scoring Elementary Schools, 4th Grade Reading 2002

Case #	RDSS_02	RDSS_03	Raw Chg Score	Residualized Chg Score
1	569	542	-27	-23.1
2	569	552	-17	-13.1
3	569	562	-7	-3.1
4	569	568	-1	2.9
5	570	549	-21	-16.9
6	570	559	-11	-6.9
7	570	562	-8	-3.9
8	570	570	0	4.2
9	571	542	-29	-24.6
10	571	557	-14	-9.6
11	571	558	-13	-8.6
12	571	559	-12	-7.6
13	571	565	-6	-1.6
14	571	567	-4	0.4
15	571	568	-3	1.4
16	571	570	-1	3.4
17	571	571	0	4.4
18	571	577	6	10.4
19	572	562	-10	-5.3
20	572	563	-9	-4.3
21	572	563	-9	-4.3
22	572	566	-6	-1.3
23	572	578	6	10.7
24	573	571	-2	2.9
25	573	572	-1	3.9
26	574	553	-21	-15.8
27	574	595	21	26.2
28	575	559	-16	-10.6
29	575	571	-4	1.5
30	576	566	-10	-4.3
31	576	567	-9	-3.3
32	576	573	-3	2.7
33	576	581	5	10.7
34	577	566	-11	-5.0
35	577	579	2	8.0
36	577	582	5	11.0
37	578	553	-25	-18.8
38	578	570	-8	-1.8
39	578	593	15	21.2
40	579	564	-15	-8.5
41	582	568	-14	-6.7
42	582	582	0	7.3
43	582	587	5	12.3
44	585	579	-6	2.0
45	585	582	-3	5.0
46	586	585	-1	7.3

47	586	591	5	13.3
48	588	585	-3	5.8
49	590	581	-9	0.3
50	595	591	-4	6.6
Mean	575.6	569.5	-6.1	-0.5
SD	6.2	12.6	9.8	10.2
Min	569	542	-29	-24.6
Max	595	595	21	26.2